Conditioning of stochastic programs

Alexander Shapiro* and Joocheol Kim
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332-0205, USA

Tito Homem-de-Mello
Department of Industrial, Welding and Systems Engineering
The Ohio State University
1971 Neil Ave., Columbus, Ohio 43210-1271, USA

Abstract

In this paper we consider stochastic programming problems where the objective function is given as an expected value function. With an optimal solution of such a (convex) problem we associate a condition number which characterizes well or ill conditioning of the problem. We show that the sample size needed to calculate the optimal solution of such problem with a given probability is approximately proportional to the condition number.

Key words: stochastic programming, Monte Carlo simulation, Large Deviations theory, ill conditioned problems.

*This work was supported, in part, by grant DMI-9713878 from the National Science Foundation.

1 Introduction

Consider the stochastic programming problem

$$\min_{x \in S} \{ f(x) := \mathbb{E}_P h(x, \omega) \}, \tag{1.1}$$

where P is a probability measure (distribution) on a sample space $(\Omega, \mathcal{F})$, S is a closed subset of $\mathbb{R}^m$ and $h : \mathbb{R}^m \times \Omega \to \mathbb{R}$ is a real valued function. We discuss in this paper ill or well conditioning of an optimal solution x_0 of the above problem (1.1). In particular we study the problem of conditioning of x_0 from the point of view of Monte Carlo sampling approximation approach. That is, suppose that an i.i.d. random sample $\omega^1, ..., \omega^N$, with the common distribution P, is generated and that the problem (1.1) is approximated by the problem

$$\min_{x \in S} \left\{ \hat{f}_N(x) := \frac{1}{N} \sum_{i=1}^{N} h(x, \omega^i) \right\}. \tag{1.2}$$

We refer to (1.1) and (1.2) as the "true" (or expected value) and the sample average approximating (SSA) problems, respectively.

In some cases the optimal solution x_0 of the true problem is stable and a relatively small sample size N is needed in order to determine x_0 with a high probability by solving the corresponding SAA problem. It is natural to say that in such cases x_0 is well conditioned, as opposed to ill conditioned problems where a much larger sample is required. Let us remark that from this point of view any problem (1.1) with multiple optimal solutions is ill conditioned.

The fact that in some cases x_0 can be calculated *exactly* (up to the computer precision) by solving the corresponding SAA problem is motivated by the following result. Suppose that the true problem (1.1) has unique optimal solution $x_0 \in S$ and let $\hat{x}_N$ be an optimal solution of the SAA problem (1.2). Under certain regularity conditions, and in particular if the distribution P is discrete, for all $\omega \in \Omega$ the function $h(\cdot, \omega)$ is piecewise linear and convex, and S is defined by linear constraints, then the event "$\hat{x}_N = x_0$" happens with probability one (w.p.1) for N large enough and, moreover, probability of that event approaches one exponentially fast as N tends to infinity. That is, there exists a constant $\beta > 0$ such that

$$\limsup_{N \to \infty} \frac{1}{N} \log [1 - P(\hat{x}_N = x_0)] \leq -\beta \tag{1.3}$$

(Shapiro and Homem-de-Mello [9]).

This is a qualitative result showing that one may not need a large sample in order to solve the true problem *exactly* with a high probability by solving the SAA problem.

The required sample size N is, of course, problem dependent and may be difficult to estimate a priori. It turns out that some problems are "well conditioned" and a small sample suffices to solve the true problem with a high probability, while other problems are "ill conditioned" and significantly a larger sample is required. One may argue that in practical applications there is no need to solve the true problem exactly. Let us remark, however, that if the true problem has a large number of optimal or nearly optimal solutions (i.e., the problem is ill conditioned), then it may be difficult to validate a calculated solution for optimality. This is because in such cases the optimal value $\hat{v}_N$ of the SAA problem gives a heavily biased estimator of the optimal value v_0 of the true problem (see Shapiro [10] and Kleywegt and Shapiro [6] for a discussion of that phenomenon).

Let us also mention that it is well known in almost every branch of numerical mathematics that large problems tend to be ill conditioned, e.g., large linear programming problems tend to be degenerate, linear regression models with a large number of predictors tend to have the multicollinearity problem, etc. Stochastic programming problems are no exceptions in this respect and large stochastic programming problems typically are ill conditioned.

In this paper we introduce a concept of the condition number associated with the optimal solution x_0 of the true problem. That condition number gives a characterization of ill (or well) conditioning of the problem from the point of view of Monte Carlo sample average approximation approach.

We use the following notation and terminology throughout the paper. By $f'(x_0, d)$ we denote the directional derivative of $f(x)$ at x_0 in the direction d. The tangent cone to a convex set S at a point $x \in S$ is denoted by $T_S(x)$, and by $S^{m-1} := \{x \in \mathbb{R}^m : \|x\| = 1\}$ we denote the unit sphere in the space $\mathbb{R}^m$. The Banach space of continuous functions $\psi : S \to \mathbb{R}$ equipped with the sup-norm is denoted by $C(S)$. By $\mathrm{Var}[X]$ we denote the variance of the random variable X.

2 Condition number

We assume that for P-almost every $\omega \in \Omega$ the function $h(\cdot, \omega)$ is real valued and convex, the feasible set S is nonempty closed and convex and is not a singleton. We also assume that the optimal solution x_0 of the true problem is *sharp*, that is

$$f'(x_0, d) > 0, \quad \forall\, d \in T_S(x_0) \setminus \{0\}. \tag{2.1}$$

Of course, the above condition (2.1) implies that the optimal solution x_0 is unique.

We say that the true problem (1.1) is *convex piecewise linear* if: (i) the set Ω is finite (and hence the distribution P is discrete), (ii) for every $\omega \in \Omega$ the function $h(\cdot, \omega)$ is convex piecewise linear, (iii) the feasible set S is convex polyhedral. For instance, two or multi-stage linear programs with recourse and discrete distributions are convex piecewise linear (see, e.g., Birge and Louveaux [1] for a discussion of two and multi-stage programming with recourse). If the true problem is convex piecewise linear, then the function $f(\cdot)$ is also convex piecewise linear. In that case the optimal solution x_0 is always sharp provided that it is unique.

Definition 2.1 *We call*

$$\kappa := \sup_{d \in T_S(x_0) \cap S^{m-1}} \frac{\mathrm{Var}\left[h'_\omega(x_0, d)\right]}{[f'(x_0, d)]^2} \tag{2.2}$$

the condition number *of the true problem* (1.1).

The above definition is motivated by the following result.

- If the true problem is convex piecewise linear and has unique optimal solution x_0, then the exponential rate (1.3) holds and the corresponding constant β is approximately equal to $[2\kappa]^{-1}$.

This means that the sample size N required to achieve a given probability of the event "$\widehat{x}_N = x_0$" is roughly proportional to the condition number κ.

Before giving a formal derivation of the above result, let us make the following remarks. Under mild regularity conditions (e.g., Shapiro and Homem-de-Mello [9]), and in particular if the true problem is convex piecewise linear, it follows that

$$\mathbb{E}_P\left[h'_\omega(x_0, d)\right] = f'(x_0, d). \tag{2.3}$$

Thus, κ can be viewed as the largest *coefficient of variation* of all random variables $h'_\omega(x_0, d),\ d \in T_S(x_0) \cap S^{m-1}$. Moreover, we have that if $\mathrm{Var}\left[h'_\omega(x_0, d)\right] = 0$, then $h'_\omega(x_0, d) = f'(x_0, d)$ for almost every $\omega \in \Omega$. If this holds for every $d \in T_S(x_0)$, then $\kappa = 0$ and in that case $\widehat{x}_N = x_0$ for any sample. Let us remark also that the condition number does not depend directly on dimensionality of the considered problem. This is because the exponential rate is faster than any polynomial rate associated with the dimension of the problem. Note, however, that as it was mentioned in the introduction large stochastic problems tend to be ill conditioned unless they have a specific, for example separable, structure.

Let us discuss now a formal justification of the above result. The following analysis is similar to Shapiro and Homem-de-Mello [9]. Consider the Banach space $Z := C(S^{m-1})$ and the set

$$F := \left\{ z \in Z : \inf_{d \in T_S(x_0) \cap S^{m-1}} z(d) \leq 0 \right\}. \tag{2.4}$$

Note that since S is not a singleton, the tangent cone $T_S(x_0)$ is not $\{0\}$, and hence the set $T_S(x_0) \cap S^{m-1}$ is not empty. Note also that the set F is closed and its interior is given by

$$\operatorname{int}(F) = \left\{ z \in Z : \inf_{d \in T_S(x_0) \cap S^{m-1}} z(d) < 0 \right\}. \tag{2.5}$$

Consider the event

$$\mathcal{E}_N := \left\{ \begin{array}{l} \text{the SAA problem (1.2) has unique optimal solution } \widehat{x}_N \\ \text{which coincides with } x_0 \end{array} \right\}. \tag{2.6}$$

The complement $\mathcal{E}_N^c$ of that event is included in the event $\{\zeta_N \in F\}$, where $\zeta_N(\cdot) := \widehat{f}'_N(x_0, \cdot)$. Note that if the true problem is convex piecewise linear, then the sample average function $\widehat{f}_N(\cdot)$ is piecewise linear and convex, and the set S coincides with the set $x_0 + T_S(x_0)$ in a neighborhood of x_0. Therefore, in the case of convex piecewise linear problems the event $\mathcal{E}_N^c$ coincides with the event $\{\zeta_N \in F\}$, and hence $P(\mathcal{E}_N^c) = P(\zeta_N \in F)$.

We assume that the Large Deviations Principle (LDP) holds for ζ_N and the set F, i.e.,

$$\begin{array}{rl} -\inf_{z \in \operatorname{int}(F)} I(z) & \leq \liminf_{N \to \infty} N^{-1} \log[P(\zeta_N \in F)] \\ & \leq \limsup_{N \to \infty} N^{-1} \log[P(\zeta_N \in F)] \leq -\inf_{z \in F} I(z). \end{array} \tag{2.7}$$

This assumption holds, for example, in case the directional derivatives $h'_\omega(x_0, d)$ are uniformly bounded with respect to d and ω. In particular, the assumption holds if the true problem is convex piecewise linear. Here $I(z)$ is the large deviations rate function, which is given by

$$I(z) := \sup_{z^* \in Z^*} \{ z^*(z) - \log M(z^*) \}, \tag{2.8}$$

and

$$M(z^*) := \int e^{z^*(z)} I\!P(dz), \quad z^* \in Z^*,$$

with $I\!P$ being the probability measure on $C(S^{m-1})$ corresponding to

$$\eta(\cdot, \omega) := h'_\omega(x_0, \cdot).$$

Now let us estimate the constant

$$\beta := \inf_{z \in F} I(z). \tag{2.9}$$

We have that

$$\beta \geq \inf_{d \in T_S(x_0) \cap S^{m-1}} I_d(0), \tag{2.10}$$

where I_d is the rate function of $\eta(d,\cdot)$. That is,

$$I_d(\alpha) := \sup_{t\in\mathbb{R}} \left[t\alpha - \Lambda_d(t)\right], \tag{2.11}$$

$M_d(t) := \mathbb{E}_P\left\{e^{t\eta(d,\omega)}\right\}$ and $\Lambda_d(t) := \log M_d(t)$. Suppose that for every $d \in T_S(x_0) \cap S^{m-1}$ the moment generating function $M_d(t)$ is finite valued in a neighborhood of zero. We have then that the moment generating function $M_d(t)$, and hence the function $\Lambda_d(t)$, are infinitely differentiable in a neighborhood of zero, and $\Lambda'_d(0) = \mathbb{E}[\eta(d,\omega)]$ and $\Lambda''_d(0) = \mathrm{Var}[\eta(d,\omega)]$. Suppose further that the variance of $\eta(d,\omega)$ is not zero, and hence $\Lambda''_d(0) > 0$.

Denote $\bar{\alpha}_d := \mathbb{E}[\eta(d,\omega)]$. Note that $\bar{\alpha}_d = f'(x_0, d)$, and hence $\bar{\alpha}_d > 0$ for any $d \in T_S(x_0) \cap S^{m-1}$ by the assumption of sharp minimum. We also have that $\bar{\alpha}_d = \Lambda'_d(0)$, and hence for $\alpha = \bar{\alpha}_d$ the maximum in the right hand side of (2.11) is attained at $t = 0$. It follows that $I_d(\bar{\alpha}_d) = -\Lambda_d(0) = 0$ and $I'_d(\bar{\alpha}_d) = 0$. Moreover, by the Implicit Function Theorem we have that

$$I''_d(\bar{\alpha}_d) = \frac{\partial^2\phi(0,\bar{\alpha}_d)}{\partial\alpha^2} - \left[\frac{\partial^2\phi(0,\bar{\alpha}_d)}{\partial t\,\partial\alpha}\right]^2 \left[\frac{\partial^2\phi(0,\bar{\alpha}_d)}{\partial t^2}\right]^{-1},$$

where $\phi(t,\alpha) := t\alpha - \Lambda_d(t)$, and hence

$$I''_d(\bar{\alpha}_d) = \frac{1}{\Lambda''_d(0)} = \frac{1}{\mathrm{Var}[\eta(d,\omega)]}.$$

Therefore, for “small" $\bar{\alpha}_d$ the second-order Taylor expansion of $I_d(\alpha)$, at $\alpha = \bar{\alpha}_d$, gives us

$$I_d(0) \approx \frac{\bar{\alpha}_d^2}{2\Lambda''_d(0)} = \frac{[\mathbb{E}\,\eta(d,\omega)]^2}{2\mathrm{Var}[\eta(d,\omega)]} = \frac{[f'(x_0,d)]^2}{2\mathrm{Var}[\eta(d,\omega)]}. \tag{2.12}$$

That is, for such d that $f'(\bar{x},d)$ is close to zero, $I_d(0)$ is approximately (up to the remainder of order $o(\bar{\alpha}_d^2)$) equal to $\frac{1}{2}[f'(\bar{x},d)]^2/\mathrm{Var}[\eta(d,\omega)]$.

The above derivations show that for ill-conditioned problems, $1/(2\kappa)$ gives approximately a lower bound for the exponential constant β. Moreover, if the true problem is convex piecewise linear, then $1/(2\kappa)$ is approximately equal to β. Indeed, in that case there exists a *finite* set $\{d_1,...,d_\ell\} \subset T_S(x_0) \cap S^{m-1}$ of directions such that if $\widehat{f}'_N(x_0,d_j) > 0$ for $j = 1,...,\ell$, then $\widehat{x}_N = x_0$, and hence the event “$\widehat{x}_N \neq x_0$" coincides with the union of the events “$\widehat{f}'_N(x_0,d_j) \leq 0$", $j = 1,...,\ell$ (Shapiro and Homem-de-Mello [9]). Probability of that union of the events is less than or equal to the sum of the probabilities of the events “$\widehat{f}'_N(x_0,d_j) \leq 0$". Also we have that probability of each event “$\widehat{f}'_N(x_0,d_j) \leq 0$" tends to zero exponentially fast with the corresponding exponential constant $I_{d_j}(0)$. Therefore, β is equal to the minimum of the numbers $I_{d_1}(0),...,I_{d_\ell}(0)$, and hence is approximately equal to $1/(2\kappa)$ if the true problem is convex piecewise linear and ill-conditioned.

3 Estimation of sample sizes

The results in the previous section provide estimates for the constant β in (1.3), which in turn yields an information on how fast the probability $P(\widehat{x}_N = x_0)$ approaches one with increase of the sample size N. Note, however, that the upper bound given by the large deviations theory for the probability $P(\mathcal{E}_N^c)$ of the event $\mathcal{E}_N$, defined in (2.6), can be quite crude for "not too large" values of N. Therefore, the above large deviations type results have more of a *qualitative* rather than a quantitative value. One might then investigate sharper estimates for $P(\mathcal{E}_N^c)$. If such estimates can be obtained, then it will be possible to compute the sample size N required to make the probability $P(\mathcal{E}_N^c)$ smaller than a specified tolerance ρ.

Let us start by discussing some general results. Consider a sequence $X_1, X_2, \dots$ of i.i.d. realizations of a (real valued) random variable X with finite mean μ and finite variance σ^2. The reader may think of X_i as the random variable $\eta(d, \omega^i) = h'_{\omega^i}(x_0, d)$, where d is a given direction and $\omega^1, \dots$ is the generated random sample. Suppose that for a given $\delta \geq 0$ we want to estimate the probability

$$p_N(\delta) := P\left(N^{-1} \sum_{i=1}^{N} X_i < \mu - \delta \right). \tag{3.13}$$

We have by the Central Limit Theorem that $N^{-1/2} \sum_{i=1}^{N} (X_i - \mu)$ converges in distribution to normal $N(0, \sigma^2)$, and hence the probability $p_N(N^{-1/2}\delta)$ tends to $\Phi(-\delta/\sigma)$ as $N \to \infty$ (here $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution).

Of course, if the random variables X_i have a normal distribution, then their average is also normally distributed, and in that case $p_N(N^{-1/2}\delta) = \Phi(-\delta/\sigma)$, or equivalently $p_N(\delta) = \Phi(-\delta\sqrt{N}/\sigma)$. Note, however, that the Central Limit Theorem does not give a justification for the asymptotics $\Phi(-\delta\sqrt{N}/\sigma)$ of $p_N(\delta)$, as $N \to \infty$, for a general distribution. We have that $\Phi(-\delta\sqrt{N}/\sigma)$ approaches zero, as $N \to \infty$, at the exponential rate $\exp(-\frac{1}{2}N\delta^2/\sigma^2)$ which can be different from the corresponding exponential rate provided by the Large Deviations theory. It is interesting to note, however, that for ill-conditioned problems (where δ^2/σ^2 is "small") the exponential rate of convergence of $p_N(\delta)$ is well approximated by the one suggested by the Central Limit Theorem (see formula (3.19) below). Let us also note that for a sample size N not "too large", $\Phi(-\delta\sqrt{N}/\sigma)$ tends to give a better approximation of $p_N(\delta)$ than the one suggested by the exact asymptotics discussed below.

We discuss now the so-called *exact asymptotics* for the probabilities $p_N(\delta)$. That theory provides an estimate $J_N(\delta)$ of $p_N(\delta)$ in the sense that $\lim_{N\to\infty} p_N(\delta)/J_N(\delta) = 1$. Let $\Lambda(\cdot)$ and $I(\cdot)$ denote the logarithmic moment generating and the rate functions of X, respectively. We assume that the moment generating function of X, and hence

$\Lambda(t)$, is *finite valued* for all t in a neighborhood $\mathcal{N}$ of zero, which implies that the mean and variance of X are finite. This assumption also implies that $\Lambda(\cdot)$ is C^∞ on $\mathcal{N}$. The following lemma shows that $\Lambda(\cdot)$ is *strictly convex* on $\mathcal{N}$.

Lemma 3.1 *Let X be a real valued random variable with positive variance such that the moment generating function of X is finite valued for all t in an open convex neighborhood $\mathcal{N}$ of zero. Then $\Lambda(\cdot)$ is strictly convex on $\mathcal{N}$.*

Proof: As it was mentioned earlier, it follows from the assumption that the moment generating function of X is finite valued for all $t \in \mathcal{N}$ that $\Lambda(\cdot)$ is C^∞ on $\mathcal{N}$. By differentiating $\Lambda(t) = \log \mathbb{E}[e^{tX}]$ we obtain, for $t \in \mathcal{N}$,

$$\Lambda''(t) \quad = \quad \frac{\mathbb{E}\left[X^2 e^{tX}\right] \mathbb{E}\left[e^{tX}\right] - \left(\mathbb{E}\left[X e^{tX}\right]\right)^2}{\left(\mathbb{E}\left[e^{tX}\right]\right)^2}.$$

Furthermore, by the Cauchy-Schwarz inequality,

$$\mathbb{E}\left[X e^{tX}\right] \;\le\; \mathbb{E}\left[|X| e^{tX}\right] \;\le\; \left(\mathbb{E}\left[X^2 e^{tX}\right]\right)^{1/2} \left(\mathbb{E}\left[e^{tX}\right]\right)^{1/2}. \tag{3.14}$$

The first inequality in (3.14) is strict if $P(X < 0) > 0$, whereas the second inequality is strict if and only if there does not exist a constant $c > 0$ such that $X^2 e^{tX} = c e^{tX}$ w.p.1 (see, e.g., Royden [7, p.121]). This of course means that the second inequality in (3.14) is strict if and only if X^2 is not a.e. constant. Moreover, since it is assumed that $\mathrm{Var}(X) > 0$, we have that X is not a.e. constant. All together this implies that at least one inequality in (3.14) is strict, and hence we have that

$$\left(\mathbb{E}\left[X e^{tX}\right]\right)^2 \;<\; \mathbb{E}\left[X^2 e^{tX}\right] \mathbb{E}\left[e^{tX}\right].$$

We obtain that $\Lambda''(t) > 0$ for all $t \in \mathcal{N}$, and hence $\Lambda(\cdot)$ is strictly convex on $\mathcal{N}$. ∎

In particular, if the set Ω is finite, then the moment generating function is finite valued for all $t \in \mathbb{R}$. In that case Lemma 3.1 shows that $\Lambda(\cdot)$ is strictly convex on $\mathbb{R}$.

Let $a \in \mathbb{R}$ be such that $\Lambda'(a) = \mu - \delta$. From Lemma 3.1 we have that $\Lambda(\cdot)$ is a strictly convex function on a neighborhood $\mathcal{N}$ of zero, and hence $\Lambda'(\cdot)$ is monotonically increasing on $\mathcal{N}$, with $\Lambda'(0) = \mu$ and $\Lambda''(0) = \sigma^2 > 0$. Therefore, for δ near zero, the solution a of the above equation exists and is unique, $a \le 0$ if $\delta \ge 0$, and $a \to 0$ as $\delta \to 0$. Moreover, $\Lambda''(a) > 0$. The estimate $J_N(\delta)$ is then given by

$$J_N(\delta) = \frac{C e^{-NI(\mu-\delta)}}{\sqrt{a^2 \Lambda''(a) 2\pi N}} \tag{3.15}$$

([2, Thm. 3.7.4]). The constant C is equal to one if X has a non-lattice distribution. Otherwise, C can be calculated as follows. Let b be the largest number such that $(X-\mu+\delta)/b$ is an integer with probability one, i.e., b is the period of the distribution of $X-\mu+\delta$. If such b does not exist, then again we take $C=1$. If such b exists (which is true for example if X is rational w.p.1 and $\mu-\delta$ is rational as well), then the constant C is given by $C=(ab)/(1-e^{-ab})$. Note that $(ab)/(1-e^{-ab})$ tends to one as $a\to 0$. Therefore, in any case we can take $C\approx 1$.

Let us now write the Taylor expansions of Λ and Λ' around zero. We have

$$\Lambda(t) \;=\; \Lambda(0)+\Lambda'(0)t+\tfrac{1}{2}\Lambda''(0)t^2+o(t^2) \;=\; \mu t+\tfrac{1}{2}\sigma^2t^2+o(t^2), \tag{3.16}$$

$$\Lambda'(t) \;=\; \Lambda'(0)+\Lambda''(0)t+o(t) \;=\; \mu+\sigma^2t+o(t), \tag{3.17}$$

and since $\Lambda'(a)=\mu-\delta$, we can approximate a (when δ, and hence a, is close to zero) by

$$a\approx -\frac{\delta}{\sigma^2}. \tag{3.18}$$

Moreover, we have that if $\Lambda'(t)=y$, then $I(y)=ty-\Lambda(t)$, and thus

$$I(\mu-\delta)=a(\mu-\delta)-\Lambda(a)=-a\delta-\tfrac{1}{2}\sigma^2a^2+o(a^2)\approx\frac{\delta^2}{2\sigma^2}.$$

Using also the approximation $\Lambda''(a)\approx\Lambda''(0)=\sigma^2$, we obtain that the estimate $J_N(\delta)$ of $p_N(\delta)$ can be approximated by

$$J_N(\delta)\;\approx\;\frac{\sigma}{\delta\sqrt{2\pi N}}\,e^{-N\delta^2/(2\sigma^2)}. \tag{3.19}$$

It is interesting to compare the above estimate of $p_N(\delta)$ with the corresponding large deviations bounds. In the present one-dimensional case the upper large deviations bound is a consequence of Chebyshev's inequality. Therefore, we have

$$p_N(\delta)\;\le\;e^{-NI(\mu-\delta)}\;\approx\;e^{-N\delta^2/(2\sigma^2)}. \tag{3.20}$$

Comparing the right hand sides of (3.19) and (3.20), we see that, while the exponential term is identical, the factor multiplying the exponential term is one in (3.20) and inversely proportional to $\sqrt{N}$ in (3.19). Therefore, the estimate $J_N(\delta)$ tends to be sharper.

Let us remark that the above estimates were computed using Taylor expansions (3.16) and (3.17). One can use, of course, higher order expansions, which will then provide more accurate estimates. This is developed in [4], to which we refer for details.

We now apply the above results to the estimation of the probability $P(\mathcal{E}_N^c)$, where $\mathcal{E}_N$ is the event defined in (2.6). In what follows, we assume that the true problem is *convex piecewise linear*. As it was mentioned in the last paragraph of section 2, in that case we have that $\widehat{x}_N$ coincides with the true solution x_0 if $\widehat{f}'_N(x_0, d_j) > 0$, $j = 1, \dots, \ell$, where $\{d_1, \dots, d_\ell\}$ is a finite set of directions independent of the sample.

For each $j = 1, \dots, \ell$, denote

$$\alpha_j := I\!E\,[h'_\omega(x_0, d_j)] = f'(x_0, d_j) \;\text{ and }\; \sigma_j^2 := \mathrm{Var}[h'_\omega(x_0, d_j)],$$

and let I_j denote the rate function of $h'_\omega(x_0, d_j)$. Then $\alpha_j > 0$ and $I_j(0)$ gives the corresponding exponential constant. Using $\delta = \alpha_j$ in (3.19) we obtain

$$P\left(\widehat{f}'_N(x_0, d_j) \le 0\right) \;\approx\; \frac{1}{\sqrt{4\pi\beta_j N}}\, e^{-\beta_j N},$$

where $\beta_j := \alpha_j^2/(2\sigma_j^2)$. We have that

$$P(\mathcal{E}_N^c) = P\left(\widehat{f}'_N(x_0, d_j) \le 0, \text{ for some } j \in \{1, ..., \ell\}\right),$$

and hence

$$P(\mathcal{E}_N^c) \le \sum_{j=1}^{\ell} P\left(\widehat{f}'_N(x_0, d_j)\right) \le 0) \approx \sum_{j=1}^{\ell} \frac{1}{\sqrt{4\pi\beta_j N}} e^{-\beta_j N} \le \frac{\ell}{\sqrt{4\pi\beta_0 N}}\; e^{-\beta_0 N}, \quad (3.21)$$

where $\beta_0 := \min_{1\le j\le \ell} \beta_j$. Note that in the present convex piecewise linear case we have that the above exponential constant β_0 is equal to $1/(2\kappa)$, where κ is the condition number defined in (2.2).

Recall that the condition $\widehat{f}'_N(x_0, d_j) > 0$, $j = 1, \dots, \ell$, is both *necessary* and *sufficient* for the occurence of the event $\mathcal{E}_N$. Therefore, for every $j \in \{1, ..., \ell\}$ we can write the following approximate lower bound for $P(\mathcal{E}_N^c)$

$$P(\mathcal{E}_N^c) \ge P(\widehat{f}'_N(x_0, d_j) \le 0) \approx \frac{1}{\sqrt{4\pi\beta_j N}}\; e^{-\beta_j N} \ge \frac{1}{\sqrt{4\pi\beta_0 N}}\; e^{-\beta_0 N}. \quad (3.22)$$

The right sides of the inequalities (3.21) and (3.22) differ from each other by the factor ℓ. This illustrates again that the condition number κ characterizes the overall rate of convergence of $P(\mathcal{E}_N^c)$ to zero.

We can now use the above results to obtain estimates of the sample size N which is needed to make $P(\mathcal{E}_N^c)$ smaller than a specified tolerance ρ. A "sufficient" condition for N can be obtained by requiring the right hand side of (3.21) to be less than ρ (the quotes are due to the fact that the inequality in (3.21) is approximate). We get

$$2\beta_0 N + \log(2\beta_0 N) \;\ge\; \log\left(\frac{\ell^2}{2\pi\rho^2}\right).$$

In order for N to satisfy the above inequality, it suffices that

$$N \;\geq\; \frac{1}{2\beta_0} \max\left\{1, \log\left(\frac{\ell^2}{2\pi\rho^2}\right)\right\} \;=\; C_1\kappa, \tag{3.23}$$

where $C_1 := \max\{1, \log(\ell^2/(2\pi\rho^2))\}$. A more accurate estimate can be obtaining by solving the nonlinear equation $z + \log z = \log(\ell^2/(2\pi\rho^2))$. By taking $z_0 = C_1$ as the initial point, this equation can be easily solved, say by Newton's method. Let C_2 denote the solution of this nonlinear equation. We can then estimate N by

$$N \;\geq\; C_2\kappa. \tag{3.24}$$

Of course, the constant (condition number) κ is unknown a priori. Note, however, that the above estimates can also be written as

$$N \;\geq\; \frac{C_i \mathrm{Var}\,[h'_\omega(x_0, d)]}{[f'(x_0, d)]^2} \qquad \text{for all } d \in T_S(x_0), \quad i = 1, 2. \tag{3.25}$$

Therefore, N can be estimated using a single direction such that $f'(x_0, d)$ is small. We discuss that in the next section.

4 Examples

We present now some examples to illustrate the ideas developed in the previous sections. Consider initially the following "median" problem. Let ω be a (one dimensional) random variable, $S := I\!R$ and $h(x, \omega) := |x - \omega|$. Suppose that ω has a discrete distribution with the odd number $r = 2k + 1$ of values equally spaced on the interval $[-1, 1]$, each having equal probability $1/r$. We have then that $x_0 = 0$ is the unique optimal solution of the true problem and for direction $d = 1$,

$$I\!E[h'(x_0, d)] \;= r^{-1} \;\text{ and }\; \mathrm{Var}[h'(x_0, d)] \;=\; 1 - r^{-2}.$$

Consequently the condition number is

$$\kappa \;=\; r^2 - 1 \;=\; 4k(k+1). \tag{4.1}$$

In that example the exact value of the exponential constant β is

$$\beta = \tfrac{1}{2}\log[r^2/(r^2-1)] \approx 1/(2r^2 - 1), \tag{4.2}$$

while the approximation $\beta \approx 1/(2\kappa)$ gives $1/(2\kappa) = 1/(2r^2 - 2)$.

Now let $\omega = (\omega_1, ..., \omega_m)$ be a random vector with independent components ω_i each having the above discrete distribution, and let $h(x, \omega) := \sum_{i=1}^m |x_i - \omega_i|$. Then

$x_0 = (0, ..., 0)$ is the optimal solution of the true problem with the same exponential constant and the condition number as in (4.1) and (4.2), respectively. This shows that this separable problem is well conditioned, and hence a small sample suffices in order to solve it exactly with high probability (see Shapiro and Homem-de-Mello [9]).

We use this example to verify the accuracy of the estimates (3.23) and (3.24) for the sample size. Let us fix $\rho = 0.05$, i.e., we wish to obtain the true optimal solution with probability 0.95. Notice that both constants C_1 and C_2 in (3.23) and (3.24) depend on the number ℓ of directions; in this separable case, we have $\ell = 2m$. Table 1 below displays the values of N obtained with (3.23) (called N_1) and (3.24) (called N_2), as well as the corresponding exact probabilities that $\widehat{x}_N = x_0$, which are given by $1 - 2P(X \geq N/2)$, where X is a binomial random variable $B(N, q)$ with $q = (r-1)/(2r)$. Those probabilities are computed for various values of m and r, as the table shows. The last column displays the ratio N_1/N_2. Notice that the number of scenarios is given by r^m. Moreover, as remarked in [9], we can see that the sample size grows quadratically with r and logarithmically with m.

In the above example, the condition number κ was known. In general, however, κ can be difficult to compute, even for simple problems, and moreover it depends on the optimal solution x_0 which, of course, is not known a priori. In the examples below we use the following procedure to estimate κ at a given optimal solution x_0: first, we generate the corresponding Monte Carlo approximation problem with sample size N_0 to obtain an approximate solution $\widehat{x}_{N_0,1}$. We then independently replicate the experiment $T-1$ more times, hence obtaining T approximate solutions $\widehat{x}_{N_0,1}, \dots, \widehat{x}_{N_0,T}$. Note that we are not interested here in the approximate objective values of the problem, but rather in the frequencies of the approximate solutions. Observe also that if the problem is ill-conditioned, then the most frequent approximate solution may not coincide with the true minimizer x_0. We exclude those $\widehat{x}_{N_0,i}$, $i = 1, \dots, T$, which coincide with x_0, and find the most frequent approximate solution from the remaining $\widehat{x}_{N_0,i}$'s. Let the chosen solution be denoted by x^*. With x_0 and x^*, we can calculate the direction $d := x^* - x_0$, and then compute the objective values at x_0 and $x_0 + d$ exactly, i.e., by enumerating all possible scenarios. Of course, these small examples allow such computations; for larger problems, one can estimate those values by large samples.

We consider now the following two numerical examples. The first example is CEP1, which was used in [9] to illustrate the exponential rate of convergence to the optimal solution. The problem was originally described in [3]. The second problem is APL1P, which was described in [5].

The CEP1 problem has 8 decision variables with 5 constraints (plus lower bound constraints) on the first stage, and 15 decision variables with 7 constraints (plus lower bound constraints) on the second stage. The random variables appear only on the right hand side of the second stage. There are 3 independent and identically distributed random variables, each taking 6 possible values with equal probability, so the sample space has size $6^3 = 216$.

For the sake of verification, we solved the problem exactly by the Benders decomposition algorithm, and obtained the true minimizer x_0 of the problem. We then solved the corresponding Monte Carlo approximating problems with sample size $N_0 = 10$ for $T = 100$ replications. After excluding the $\widehat{x}_{N_0,i}$'s that coincide with the true minimizer x_0 of the problem, we chose the most frequently obtained optimal solution, x^*, among the remaining $\widehat{x}_{N_0,j}$'s. We then calculate $f'(x_0, d)$ and $\mathrm{Var}[h'_\omega(x_0, d)]$ for the direction $d := x^* - x_0$. Note that this direction is likely to have the largest condition number. Table 2 below displays the results. The table also displays the value of N estimated with (3.24) that guarantees that the optimal solution will be obtained with probability at least 0.95. Note that this requires an estimate for ℓ. In this case we chose $\ell = 1$ due to the small number of decision variables. Note also that the estimate obtained for N ($N \geq 54$) is in agreement with the results obtained in [9] — in that paper it was verified computationally that a sample size equal to 50 yields the optimal solution with probability 0.97.

The APL1P example is an electric power capacity expansion problem on a transportation network. The problem has two decision variables with 2 constraints (plus lower bound constraints) on the first stage, and 9 decision variables with 5 constraints (plus lower bound constraints) on the second stage. The random variables appear on both the right hand side and the technology matrix of the second stage. There are 5 independent random variables. The number of realizations per random variables yields a total of $4 \times 5 \times 4 \times 4 \times 4 = 1280$ scenarios. To estimate κ, we used the same procedure outlined above, with sample size $N_0 = 200$ and $T = 100$ replications. As with the CEP1 problem, table 2 below displays the value of N estimated with (3.24) that guarantees that the optimal solution will be obtained with probability at least 0.95. Note that the estimate obtained for N is larger than the total number of scenarios. That happened since the problem is small and ill conditioned. Of course, it makes sense to use Monte Carlo sampling techniques only for problems with a very large number of scenarios, so this example is given for illustration purposes only.

References

[1] J.R. Birge and F. Louveaux, *Introduction to Stochastic Programming*, Springer, New York, 1997.

[2] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd. ed., Springer-Verlag, New York, 1998.

[3] J.L. Higle and S. Sen, "Finite master programs in regularized stochastic decomposition", *Mathematical Programming*, 67 (1994), 143-168.

[4] T. Homem-de-Mello, "Monte Carlo methods for discrete stochastic optimization", preprint, 2000.

[5] G. Infanger, "Monte Carlo (importance) sampling within a Benders decomposition algorithm for stochastic linear programs," *Annals of Operations Research*, 39, 1992, pp. 69-95.

[6] A. Kleywegt and A. Shapiro, "The Sample Average Approximation Method for Stochastic Discrete Optimization", preprint. Available at: *Stochastic Programming E-Print Series*, `http://dochost.rz.hu-berlin.de/speps/`

[7] H. Royden, *Real Analysis*, 3rd. ed., Macmillan, New York, 1988.

[8] R.Y. Rubinstein and A. Shapiro, *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*. Wiley, New York, NY, 1993.

[9] A. Shapiro and T. Homem-de-Mello, "On the rate of convergence of optimal solutions of Monte Carlo approximations of stochastic programs", *SIAM J. Optimization*, to appear.

[10] A. Shapiro "Stochastic programming by Monte Carlo simulation methods", *Journal Statistical Planning and Inference*, to appear.

r	m	N_1	p_{N_1}	N_2	p_{N_2}	N_1/N_2
5	5	211	0.984	165	0.954	1.279
5	10	244	0.980	194	0.941	1.258
5	100	355	0.986	294	0.937	1.207
5	500	432	0.984	366	0.934	1.180
5	1000	465	0.987	398	0.934	1.168
11	5	1052	0.983	821	0.956	1.281
11	10	1218	0.984	968	0.950	1.258
11	100	1771	0.987	1470	0.949	1.205
11	500	2157	0.988	1830	0.948	1.179
11	1000	2323	0.989	1986	0.947	1.170
21	5	3854	0.984	3009	0.956	1.281
21	10	4464	0.985	3546	0.953	1.259
21	100	6491	0.988	5388	0.952	1.205
21	500	7907	0.989	6708	0.951	1.179
21	1000	8517	0.989	7282	0.951	1.170
31	5	8409	0.985	6564	0.955	1.281
31	10	9740	0.985	7737	0.956	1.259
31	100	14161	0.988	11756	0.953	1.205
31	500	17251	0.989	14636	0.952	1.179
31	1000	18582	0.989	15888	0.952	1.170
51	5	22773	0.985	17775	0.956	1.281
51	10	26378	0.985	20952	0.955	1.259
51	100	38351	0.988	31838	0.953	1.205
51	500	46720	0.989	39637	0.954	1.179
51	1000	50325	0.989	43028	0.953	1.170

Table 1: Estimated sample sizes to attain probability 0.95 and exact probabilities $P(\widehat{x}_N = x_0)$ for the median problem

	CEP1	APL1P
$f'(x_0, d)$	7.61	0.06
$\mathrm{Var}[h'_\omega(x_0, d)]$	1010.16	4.02
$\kappa \geq$	17.45	1105.58
$N \geq$	54	3363

Table 2: Condition number and sample size estimates for the CEP1 and APL1P problems